Ecological Informatics 7 (2012) 19–29

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Hierarchical classification of diatom images using ensembles of predictive clustering trees

Ivica Dimitrovski [a,*], Dragi Kocev [b], Suzana Loskovska [a], Sašo Džeroski [b]

[a] Department of Computer Science and Computer Engineering, Faculty of Electrical Engineering and Information Technologies, Rugjer Boshkovik bb, 1000 Skopje, Republic of Macedonia
[b] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

A R T I C L E I N F O

Article history:
Received 5 July 2011
Received in revised form 4 September 2011
Accepted 5 September 2011
Available online 10 September 2011

Keywords:
Diatoms
Automatic image annotation
Hierarchical classification
Predictive clustering trees
Feature extraction from images

A B S T R A C T

This paper presents a hierarchical multi-label classification (HMC) system for diatom image classification. HMC is a variant of classification where an instance may belong to multiple classes at the same time and these classes/labels are organized in a hierarchy. Our approach to HMC exploits the classification hierarchy by building a single predictive clustering tree (PCT) that can simultaneously predict all different levels in the hierarchy of taxonomic ranks: genus, species, variety, and form. Hence, PCTs are very efficient: a single classifier is valid for the hierarchical classification scheme as a whole. To improve the predictive performance of the PCTs, we construct ensembles of PCTs. We evaluate our system on the ADIAC database of diatom images. We apply several feature extraction techniques that can be used in the context of diatom images. Moreover, we investigate whether the combination of these techniques increases predictive performance. The results show that ensembles of PCTs have better predictive performance and are more efficient than SVMs. Furthermore, the proposed system outperforms the most widely used approaches for image annotation. Finally, we demonstrate how the system can be used by taxonomists to annotate new diatom images.

1. Introduction

Diatoms are a large and ecologically important group of unicellular or colonial organisms (algae). They are characterized by their highly patterned cell wall composed mainly of hydrated amorphous silica. The cell wall can be divided into two halves. Each half of the cell wall consists of a valve and a number of girdle bands. One half is slightly larger than the other and overlaps it. Together, the halves make a cylinder, with the two valves at the ends. The cross section of the cylinder, and hence the outline of the valve, varies greatly in shape between species and genera. This, together with the pattern of pores and other markings on the valve, provides the information needed for species classification. Fig. 1 depicts three example images of diatoms.

In the variety of uses of diatoms, such as water quality monitoring, paleoecology and forensics, microscope slides must be first scanned for diatoms: if diatoms are present, they need to be classified. Most classifications are done using classification keys and/or comparing specimens using slides, photographs or drawings of diatoms in books and atlases (Stoermer and Smol, 2004). This is not a trivial task, taking into consideration that taxonomists estimate that there may be 200,000 different diatom species, half of them still undiscovered, and many of these extremely hard to distinguish on the basis of morphology (du Buf and Bayer, 2002). Furthermore, this is very tedious and repetitive work, thus any degree of automation can greatly help.

Having this in mind, we propose a system for automatic diatom classification. This system consists of the two standard parts of image annotation systems: image processing (feature extraction from images) and image classification. The image processing part converts an image to a set of numerical features that are extracted directly from the image pixels. The second part, image classification, labels and groups the images. The labels can be organized in a hierarchy and an image can be labeled with more than one label (can belong to more than one group).

For the image processing part, we have implemented two feature extraction techniques that are most commonly used in this context. The first technique produces descriptors (called Fourier descriptors) that contain information concerning the properties of the valve outline. The descriptors from the second technique, called Scale Invariant Feature Transform-SIFT histograms, contain information about the ornamentation of the valve face. We believe that the diatom images can be appropriately described with the combination of these two techniques.

Considering the image classification part, we will use the recently proposed method of building ensembles of PCTs, in particular, bagging and random forests of PCTs (Kocev et al.,2007; Kocev, 2011). We directly compare the predictive performance and the efficiency of the ensembles of PCTs to the one of SVMs trained with Gaussian kernels — the most widely used classifiers used in image annotation.

* Corresponding author. Tel.: +389 2 3099156.
E-mail addresses: ivicad@feit.ukim.edu.mk (I. Dimitrovski), Dragi.Kocev@ijs.si (D. Kocev), suze@feit.ukim.edu.mk (S. Loskovska), Saso.Dzeroski@ijs.si (S. Džeroski).

1574-9541/$ – see front matter
doi:10.1016/j.ecoinf.2011.09.001

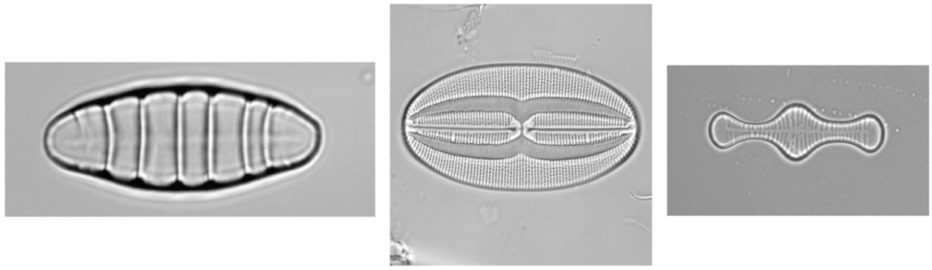

Fig. 1. Example images of diatoms. From left to right: *Diatoma mesodon*, *Fallacia sp.5* and *Tabellaria flocculosa*.

Moreover, we contrast the predictive performance of the proposed approach to the best reported results (du Buf and Bayer, 2002) on the used database of diatom images (ADIAC, 2011).

The goal of the complete system is to assist a taxonomist in identifying a wide range of different diatoms. To this end, we develop a web-based interface to the proposed system than can help taxonomists with the identification of diatom taxa in images. The user (i.e., taxonomist) loads an image to the system by using the interface. After that, the system recommends an annotation for the image, accompanied with a probability value for the prediction. The taxonomist can then select the proposed annotation and browse through the images from the same species to further check the validity of the annotation.

The remainder of the paper is organized as follows. In Section 2, we present the related work. Section 3 describes the system for annotation of diatom images, namely, image segmentation, techniques we use for feature extraction from images, and predictive clustering trees and their use for HMC. In Section 4, we explain the experimental setup. The obtained results and a discussion thereof are given in Section 5. Section 6 concludes the paper and points out some directions for further work.

2. Background and related work

The process of automatic diatom classification consists of three phases (du Buf and Bayer, 2002): image segmentation, feature extraction and image classification. The goal of image segmentation is to locate and obtain the contour of the diatom. Then, using these segmented images and extracted contours, the feature extraction algorithms generate image descriptors. At the end, machine learning algorithms are used to train a classifier that will perform the classification for previously unseen diatom images (e.g., provide the taxonomic rank). Here, we shortly describe each phase and the algorithms which are usually used in each of them.

2.1. Image segmentation

An ideal diatom image depicts only a single diatom shell. However, in reality, diatoms may lay on top of each other or very close to each other, the image may not be in proper focus, dust specks and background texture may be visible in some images etc. Fig. 3 (first row) shows diatom images that contain some of the aforementioned anomalies. Because of this, one needs to perform image segmentation before extracting features from the images.

The problem of image segmentation, i.e., contour extraction, of gray-scale diatom images can be solved mainly by applying four methods: threshold-based, boundary-based, region-based and hybrid methods (Jalba et al., 2004). *Threshold methods* assume that all pixels with gray-level values within a certain range belong to one class. They do not use any spatial information of the image, are sensitive to noise, and do not cope well with blurred edges. The *boundary-based methods* are local filtering techniques, such as edge detectors or active contour methods. Edge detectors usually cannot ensure continuous edge-detection and an edge-linking step must be used to produce closed contours. In contrast, active contour methods automatically produce closed contours and usually provide better edge localization, but are sensitive to noise and require an initialization step that is hard to automate. *Region-based methods* assume that neighboring pixels within the same region have similar values. Their main advantage is that they use and adapt the statistics inside the region, but they generate small holes and irregular boundaries. *Hybrid techniques* combine both boundary and region criteria. All in all, there is a variety of approaches that one can choose for the problem at hand. In our system, we are using marker-controlled watershed segmentation which has already been successfully applied for diatom image segmentation (Jalba et al., 2004).

2.2. Feature extraction

Once the segmentation and contour extraction are completed, different feature extraction techniques can be employed on the diatom images (Westenberg and Roerdink, 2002). The diatoms can be primarily distinguished by evaluating properties of the valve's outline. The contour features measure the symmetry, global and local shape characteristics, as well as geometric properties, such as length and width of the diatoms (Ciobanu and du Buf, 2002; Fischer and Bunke, 2002; Loke and du Buf, 2002).

An important characteristic of diatoms is also the ornamentation of the valve face, which is a specific type of texture (Wilkinson et al., 2002). There are several known visual descriptors able to measure these texture properties: features derived from gray level co-occurrence matrices, Gabor wavelets (Santos and du Buf, 2002), scale invariant feature transform (SIFT) (Lowe, 2004) and local binary patterns (LBP) (Ojala et al., 2002). To summarize, these features capture several aspects of an image. Depending on the application, one can choose to use some specific feature extraction technique or to combine several of them into a single, more complex set of features.

2.3. Image classification

The last phase of an automatic classification system is classification. In this phase, a machine learning algorithm is first employed to construct a classifier using the features extracted in the previous two stages and the annotations/labels of the images (taxonomic ranks). Then, the obtained classifier maps the images of unidentified specimens to annotations from the set observed during training, i.e., provides annotations for previously unseen images. In the context of diatom image classification, the most typically used classifiers are neural networks, naïve Bayes, support vector machines (SVMs) and decision trees.

Santos and Du Buf (Santos and du Buf, 2002) use a fully-connected neural network classifier with one hidden layer. The number of input units equals the number of features. The hidden layer has an equal number of units as the input layer, and the output layer has as many units as there are classes. The neural network is trained until the error rate on a validation set reaches a local minimum.